Análise e predição de incidência de casos de malária no tempo e no espaço utilizando modelos deep learning Resumo Estendido - CTDSI/CTCCSI 2021

Matheus Félix Xavier Barboza[1], Vanderson de Souza Sampaio (co-orientador)[2,3], Patricia Takako Endo (orientadora)[1]

[1]Universidade de Pernambuco (UPE)
Pernambuco – Brasil

[2]Fundação de Vigilância em Saúde do Amazonas (FVS-AM)
Amazonas – Brasil

[3]Fundação de Medicina Tropical Dr. Heitor Vieira Dourado (FMT-HVD)
Amazonas – Brasil

matheus.barboza@upe.br, vandersons@gmail.com, patricia.endo@upe.br

Resumo. *A malária é uma doença com risco de vida evitável e curável, mas houve mais de 228 milhões de casos de malária e 405.000 mortes por malária em 2018. Enquanto mais de 42 milhões de brasileiros estão sob risco de malária, 99% de todos os casos de malária no Brasil estão localizados dentro ou ao redor da floresta amazônica. Apesar do declínio de casos e mortes, a malária continua sendo um grande problema de saúde pública no Brasil. Em resposta a pedidos de novas pesquisas sobre estratégias de eliminação da malária para atender às condições locais, este artigo propõe modelos de machine learning e deep learning para prever a probabilidade de casos de malária no Estado do Amazonas. Usando um conjunto de dados de aproximadamente 6 milhões de registros, avaliamos os modelos Random Forest, LSTM e GRU e comparamos desempenho por área geográfica usando a classificação de regionais de saúde do Estado do Amazonas e clusters através do algoritmo k-means. Os resultados sugerem que todos os modelos têm uma precisão satisfatória e forte potencial para prever novos casos de malária na região.*

Abstract. *Malaria is a preventable and curable life-threatening disease yet there was over 228 million cases of Malaria and 405,000 deaths from Malaria in 2018. Over 42 million Brazilians are at risk from Malaria; 99% percent of all malaria cases in Brazil are located in or around the Amazon rainforest. Despite declining cases and deaths, Malaria remains major public health issue in Brazil. In response to calls for novel research for the adaptation of Malaria mitigation and eradication strategies to suit local conditions, this paper proposes machine learning and deep learning models to predict the probability of Malaria cases in the State of Amazonas. Using a data set of approximately 6 million records, we evaluate Random Forest, LSTM and GRU models and we compared performance by geographic area using the classification of health regions of the State of Amazonas and clusters using the k-means algorithm. The results suggest that all models have satisfactory precision and strong potential to predict new cases of malaria in the region.*

1. Introdução

A malária é uma doença transmitida por mosquitos, especificamente pelas fêmeas do gênero *Anopheles*. No Brasil, é possível notar um impacto significativo da malária, pois, ainda que as taxas se encontrassem estáveis nos anos de 2014 a 2016 [WHO, 2018]. Em 2017, houve um aumento do número de casos, com um total estimado de 218 mil [WHO, 2018]. Estudos mostram que cerca de 99% dos casos da doença no Brasil estão concentrados na área da Floresta Amazônica [Carlos et al., 2019], principalmente nos Estados do Acre e do Amazonas. Assim, dada a atual literatura, nosso trabalho apresenta duas principais contribuições: (a) a primeira contribuição é a proposição e a avaliação de modelos de *deep learning* para prever a ocorrência de malária no Estado do Amazonas, sendo um dos primeiros estudos que aplicam modelos de *deep learning* sobre dados de malária no Brasil, além de fazer uso de um conjunto de dados substancial; e (b) a segunda contribuição é a metodologia aplicada para a criação de *cluster* de municípios, de acordo com suas semelhanças estatísticas de casos de malária, para auxiliar na predição dos casos da doença.

2. Materiais e métodos

Neste trabalho, dois modelos de predição são propostos para estimar o número de casos de malária: *Long-short Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU). Os modelos são avaliados tendo como base um modelo tradicional de *machine learning*, o *Random Forest*. Para tanto, este trabalho utilizou *(a)* a base de dados do SIVEP-MALÁRIA, e *(b)* a técnica de *clusterização k-means* para agrupar os municípios com características estatísticas similares baseado nos dados de ocorrência de malária.

2.1. Base de Dados

Neste trabalho, foi utilizado um conjunto de dados relacionados a casos de malária de janeiro de 2003 a dezembro de 2018 referentes ao Estado do Amazonas, com aproximadamente 6 milhões de registros, provenintes do SIVEP-MALÁRIA. Entre 2003 e 2007, o número de casos foi superior aos anos seguintes, atingindo cerca de 30 mil ocorrências em julho de 2005. Embora a incidência da malária tenha diminuído, ainda permanece relativamente alta entre alguns municípios.

2.2. Clusterização e Regionais de Saúde

A clusterização é uma técnica que divide as amostras de um conjunto de dados em grupos por elementos com características semelhantes [Kopec, 2019]. O algoritmo *k-means* é um dos métodos mais conhecidos para executar o agrupamento de dados, que consiste em particionar um número predefinido de clusters, k, usando uma classificação não supervisionada. Para realizar a análise dos modelos propostos, além de considerar os clusters de municípios, também foi levada em consideração as Regionais de Saúde do Amazonas, composta por nove regionais contendo municípios de acordo com variáveis como taxa populacional, porcentagem populacional relativa, possibilidades de acesso e localização geográfica [da Silva et al., 2018].

3. Modelos de previsão

Os modelos *deep learning*, LSTM e GRU, propostos para prever a ocorrência de malária possuem a mesma arquitetura, composta por duas camadas (LSTM ou GRU), ambas com

cinquenta unidades por camada, seguidas por uma camada totalmente conectada com uma unidade que fornece a previsão da malária como saída. Após cada camada recorrente (LSTM e GRU), usamos a técnica de *dropout* com uma probabilidade igual a 20%. Os parâmetros (como o número de camadas e unidades) foram escolhidos empiricamente.

Usamos o método de validação para realizar os experimentos. Selecionamos 80% dos dados históricos para o treinamento do modelo e 20% para o teste. Executamos cada técnica 10 vezes, depois coletamos a média da do erro quadrático médio quadrático (RMSE, do inglês *root-mean-square error*) para avaliação dos modelos.

4. Resultados e Análise estatística

Os modelos de *deep learning* e *machine learning* obtiveram resultados de predição de ocorrência de malária bastante semelhantes quando aplicados nos cenários de clusters compostos pelo algoritmo *k-means* e quando considerando as Regionais de Saúde. As médias de RMSE em ambos os cenários foram bastante próximas, com exceção do cluster 5, que se distanciou dos demais, obtendo os piores resultados nos três modelos. Outra semelhança entre os dois cenários foi a predominância de melhores resultados do modelo GRU, no qual sete clusters e oito Regionais de Saúde se destacaram com esta técnica.

5. Conclusão

De uma perspectiva de *deep learning*, nossos resultados são consistentes com estudos existentes [Chung et al., 2014]. O modelo GRU obtém melhores resultados em mais da metade dos testes realizados, mostrando que mesmo com valores muito semelhantes, ainda supera o LSTM [Chung et al., 2014]. Apesar do desempenho comparável do LSTM e GRU, o último geralmente apresenta tempo de treinamento significativamente mais curto [Chung et al., 2014, Jozefowicz et al., 2015], o que pode ser vantajoso na prática. Os resultados sugerem que todos os modelos têm precisão satisfatória e forte potencial para prever novos casos nos municípios.

Referências

Carlos, B. C., Rona, L. D., Christophides, G. K., and Souza-Neto, J. A. (2019). A comprehensive analysis of malaria transmission in brazil. *Pathogens and global health*, 113(1):1–13.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

da Silva, J. C., Monteiro, A. C. P., and de Souza Fonseca, S. (2018). Plano estadual de educação permanente em saúde do amazonas – brasil - 2019 - 2020.

Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning*, pages 2342–2350.

Kopec, D. (2019). *Classic Computer Science Problems in Python*. Manning Publications Co.

WHO (2018). Malaria profile - brazil 2018.